A NOTE ON NON-PARAMETRIC BAYESIAN ESTIMATION FOR POISSON POINT PROCESSES

SHOTA GUGUSHVILI AND PETER SPREIJ

Abstract. We derive the posterior contraction rate for non-parametric Bayesian estimation of the intensity function of a Poisson point process.

1. Introduction

Poisson point processes (see e.g. Kingman (1993)) are among the basic modelling tools in areas as different as astronomy, biology, image analysis, reliability theory, medicine, physics, and others. A Poisson point process X on the space $\mathcal{X} = [0,1]^d$ (this is good enough for our purposes) with the Borel σ-field $\mathcal{B}(\mathcal{X})$ of its subsets is a random integer-valued measure on $\mathcal{X}$ (we assume the underlying probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ in the background), such that

(i) for any disjoint subsets $B_1, B_2, \ldots, B_m \in \mathcal{B}(\mathcal{X})$, the random variables $X(B_1), X(B_2), \ldots, X(B_m)$ are independent, and
(ii) for any $B \in \mathcal{B}(\mathcal{X})$, the random variable $X(B)$ is Poisson distributed with parameter $\Lambda(B)$, where Λ is a finite measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, called the intensity measure of the process X.

Intuitively, the process X can be thought of as random scattering of points in $\mathcal{X}$, where scattering occurs in a special way determined by properties (i)–(ii) above.

In practical applications knowledge of the intensity Λ is of importance. The latter typically cannot be assumed known beforehand and has to be estimated based on the observational data on the process X. A popular assumption in the literature (see e.g. the references on p. 263 in Kutoyants (1998)) is that one has independent observations $X_1, \ldots, X_n$ on the process X over $\mathcal{X}$ at his disposal, on basis of which an estimator of Λ has to be constructed. We will denote for brevity $X^{(n)} = (X_1, X_2, \ldots, X_n)$. In case Λ is absolutely continuous with respect to some dominating measure and has a density λ, one might also be interested in estimation of λ. We will assume that Λ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and will call λ the intensity function.

From now on we concentrate on estimation of the intensity function. Two broad approaches to estimation of λ, parametric and non-parametric, can be discerned in the literature. In the parametric approach, one assumes that the unknown intensity function λ can be parametrised by a finite-dimensional parameter θ (where, for instance, θ ranges in some subset Θ of $\mathbb{R}^p$), so that $\lambda = \lambda_\theta$, and the corresponding

Date: April 30, 2013.

2000 *Mathematics Subject Classification.* Primary: 62G20, Secondary: 62M30.

Key words and phrases. Intensity function; Non-parametric Bayesian estimation; Poisson point process; Posterior contractaion rate.

The research of the first author was supported by The Netherlands Organisation for Scientific Research (NWO).

statistical experiment generated by $X^{(n)}$ is denoted by $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n), \{\mathbb{P}_\theta^{(n)}, \theta \in \Theta\})$. The goal is to estimate the 'true' parameter θ_0 on the basis of the sample $X^{(n)}$. In the non-parametric approach to estimation of λ no such assumptions are made. Instead, one assumes, for instance, that λ belongs to some class Θ of functions possessing given smoothness properties (the statistical experiment generated by $X^{(n)}$ is $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n), \{\mathbb{P}_\lambda^{(n)}, \lambda \in \Theta\})$), and the goal is to estimate the 'true' intensity function λ_0. See e.g. Kutoyants (1998) for additional information on statistical inference for Poisson point processes from the point of view of asymptotic statistical theory. Computational approaches are studied e.g. in Møller and Waagepetersen (2004) and are reviewed in Møller and Waagepetersen (2007).

In this note we are interested in non-parametric estimation of the intensity function λ_0. A kernel-type estimator of λ_0 has been studied in detail in Section 6.2 in Kutoyants (1998), see also p. 263 there for further references. In particular, it is shown in Kutoyants (1998) that this estimator is optimal in the minimax sense over the class of β-Hölder-regular intensity functions.

Here we will take an alternative, non-parametric Bayesian approach to estimation of λ_0, but will analyse it from the frequentist point of view. In the Bayesian approach to estimation of λ_0 one puts a prior Π on λ_0, which might be thought of as reflecting one's prior knowledge or belief in λ_0. In more formal terms this is a measure Π defined on the parameter set Θ equipped with some σ-field $\sigma(\Theta)$, and one assumes that $\lambda_0 \in \Theta$. The set Θ equipped with a certain σ-field $\sigma(\Theta)$ is a set of finite-valued functions defined on $[0,1]^d$, which we for technical reasons assume to be uniformly bounded away from zero. Then by Theorem 1.3 in Kutoyants (1998), for any $\lambda \in \Theta$, the law $\mathbb{P}_\lambda$ of X under the parameter value λ admits a density p_λ with respect to the measure $\mathbb{P}_{\text{sp}}$ induced by a standard Poisson point process with intensity measure $\Lambda_{\text{sp}}(\mathrm{d}x) = \mathrm{d}x$. This density is given by

$$p_\lambda(\xi) = \exp\left(\int_{[0,1]^d} \log \lambda(x) \xi(\mathrm{d}x) - \int_{[0,1]^d} [\lambda(x) - 1] \mathrm{d}x \right),$$

where $\xi = \sum_{i=1}^m \delta_{x_i}$ is a realisation of X (here δ_{x_i} denotes the Dirac measure at x_i) and

$$\int_{[0,1]^d} \log \lambda(x) \xi(\mathrm{d}x) = \sum_{i=1}^m \log(\lambda(x_i)).$$

Using independence of X_i's, it follows that the likelihood $L_\lambda(X^{(n)})$ for $X^{(n)}$ can be written as

$$L_\lambda(X^{(n)}) = \prod_{i=1}^n \exp\left(\int_{[0,1]^d} \log \lambda(x) X_i(\mathrm{d}x) - \int_{[0,1]^d} [\lambda(x) - 1] \mathrm{d}x \right). \tag{1}$$

Assuming joint measurability of $p_\lambda(\xi)$ in (ξ, λ), so that the integrals below make sense, Bayes' formula gives the posterior measure of any measurable set $A \in \sigma(\Theta)$ through

$$\Pi(A|X^{(n)}) = \frac{\int_A L_\lambda(X^{(n)}) \mathrm{d}\Pi(\lambda)}{\int_\Theta L_\lambda(X^{(n)}) \mathrm{d}\Pi(\lambda)}.$$

Transition from the prior to the posterior can be thought of as updating our prior opinion on λ_0 upon seeing the data $X^{(n)}$.

We will study the rate of convergence of the posterior distribution $\Pi(\cdot|X^{(n)})$ under $\mathbb{P}_{\lambda_0}^{(n)}$, where $\mathbb{P}_{\lambda_0}^{(n)}$ denotes the law of $X^{(n)}$ under the true parameter value λ_0.

The idea, informally speaking, is that with the sample size n growing indefinitely, the Bayesian approach should be able to recognise the true λ_0 with increasing accuracy. This can be formalised by requiring, for instance, that for any fixed neighbourhood A of λ_0, $\Pi(A^c|X^{(n)}) \to 0$ in $\mathbb{P}_{\lambda_0}^{(n)}$-probability, or, in words, by requiring that with the Bayesian approach with a prior Π, most of the posterior mass must eventually concentrate around the true parameter value λ_0. More generally, one might take a sequence of shrinking neighbourhoods A_n of λ_0 and ask what is the fastest rate, at which the neighbourhoods A_n can shrink, while still capturing most of the posterior mass (the precise definition will be given below). The case for such an approach to the study of non-parametric Bayesian techniques is made e.g. in Diaconis and Freedman (1986), while several recent references dealing with establishing posterior convergence rates under broad conditions in various statistical settings are Ghosal et al. (2000), Ghosal and van der Vaart (2001) and van der Vaart and van Zanten (2008a). The rate, at which the neighbourhoods A_n shrink, can be thought of as an analogue of the convergence rate of a frequentist estimator. The analogy can be made precise in the sense that contraction of the posterior distribution at a certain rate implies existence of a Bayes point estimate with the same convergence rate (in the frequentist sense); see e.g. Theorem 2.5 in Ghosal et al. (2000) and the discussion on pp. 506–507 there.

The rest of the paper is organised as follows: in the next section we state the problem we are interested in in greater detail and provide a general result on the posterior contraction rate in our problem with the prior based on a transformation of a Gaussian processes. In Section 3 we consider a concrete example of the prior and compute the posterior contraction rate explicitly. Finally, Appendix A contains the proof of the technical lemma used in the proof of our main theorem. Computational aspects of the non-parametric Bayesian approach to the intensity function estimation lie outside the scope of this note. Instead we refer to Heikkinen and Arjas (1998) for one specific implementation.

2. Main result

In order to study the contraction rate of the posterior distribution in our setting, we first need to specify the suitable neighbourhoods A_n of λ_0, for which this will be done. The Hellinger distance $h(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_1})$ between two probability laws $\mathbb{P}_{\lambda_1}$ and $\mathbb{P}_{\lambda_1}$ is defined as

$$\begin{aligned} h(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}) &= \left\{ \int (\mathrm{d}\mathbb{P}_{\lambda_1}^{1/2} - \mathrm{d}\mathbb{P}_{\lambda_2}^{1/2})^2 \right\}^{1/2} \\ &= \left\{ \int (p_{\lambda_1}^{1/2} - p_{\lambda_2}^{1/2})^2 \mathrm{d}\mathbb{P}_{\mathrm{sp}} \right\}^{1/2}. \end{aligned}$$

Here, as in Section 1, we assume that λ_i's are bounded away from zero and infinity, which yields in particular the second equality in the above display. The Hellinger distance is one of the popular discrepancy measures between two probability laws. The Hellinger distance can also be used to define the pseudo-distance $\rho(\lambda_1, \lambda_2)$ between parameters λ_1 and λ_2 by setting

$$\rho(\lambda_1, \lambda_2) = h(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}).$$

Thus λ_1 and λ_2 are close to each other if the corresponding laws $\mathbb{P}_{\lambda_1}$ and $\mathbb{P}_{\lambda_1}$ are in Hellinger distance. We also introduce two further discrepancy measures: the

Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})$ between two probability laws $\mathbb{P}_{\lambda_1}$ and $\mathbb{P}_{\lambda_2}$ is defined as

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2}) &= \int \log\left(\frac{\mathrm{d}\mathbb{P}_{\lambda_1}}{\mathrm{d}\mathbb{P}_{\lambda_2}}\right)\mathrm{d}\mathbb{P}_{\lambda_1}\\ &= \int p_{\lambda_1}\log\left(\frac{p_{\lambda_1}}{p_{\lambda_2}}\right)\mathrm{d}\mathbb{P}_{\mathrm{sp}},\end{aligned}$$

while the discrepancy V is defined through

$$\begin{aligned}\mathrm{V}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2}) &= \int\left(\log\left(\frac{\mathrm{d}\mathbb{P}_{\lambda_1}}{\mathrm{d}\mathbb{P}_{\lambda_2}}\right)-\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})\right)^2\mathrm{d}\mathbb{P}_{\lambda_1}\\ &= \int p_{\lambda_1}\left(\log\left(\frac{p_{\lambda_1}}{p_{\lambda_2}}\right)-\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})\right)^2\mathrm{d}\mathbb{P}_{\mathrm{sp}}.\end{aligned}$$

This can be thought of as the Kullback-Leibler 'variance'. Both quantities are well-defined, because under our standing assumption that λ_1 and λ_2 are bounded away from zero, the corresponding laws $\mathbb{P}_{\lambda_1}$ and $\mathbb{P}_{\lambda_2}$ are equivalent.

We will derive the posterior convergence rate by taking the neighbourhoods A_n of λ_0 to be balls of appropriate radii in the pseudo-distance ρ, see below. This is a reasonable choice, see e.g. Ghosal et al. (2000).

We need to specify the prior Π. Priors based on stochastic processes are widely used in Bayesian statistics. In particular, priors based on Gaussian processes are a popular choice both in the statistics and machine learning communities, see e.g. Rasmussen and Williams (2006), as well as van der Vaart and van Zanten (2008a) for additional references. For our purposes, a zero-mean Gaussian process $W = (W_x)_{x\in\mathcal{X}}$ is a collection of random variables W_x indexed by $\mathcal{X}$ and defined on the common probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$, such that the finite-dimensional distributions of W are zero-mean multivariate normal distributions. The latter are determined by the covariance function $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, defined by

$$K(x,y)=\widetilde{\mathbb{E}}\,[W_xW_y],\quad x,y\in\mathcal{X},$$

where $\widetilde{\mathbb{E}}$ denotes the expectation with respect to the measure $\widetilde{\mathbb{P}}$. For all the necessary definitions and properties of Gaussian processes with a view towards applications in non-parametric Bayesian statistics that are used in this work, see van der Vaart and van Zanten (2008a) and van der Vaart and van Zanten (2008b).

Assume that W is a zero-mean Gaussian process with bounded sample paths $x\mapsto W_x$ and let $\kappa>0$ be a fixed constant. Define the process $Z^{(W)}=\left(Z_x^{(W)}\right)_{x\in\mathcal{X}}$ through

$$Z_x^{(W)}=\kappa+|W_x|,\quad x\in\mathcal{X}. \tag{2}$$

Realisations of W will be denoted by lowercase letters, such as w and v. The corresponding realisations of $Z^{(W)}$ will be denoted by $z^{(w)}$ and $z^{(v)}$. Our prior Π will be the law of the process $Z^{(W)}$, which implicitly defines our parameter set Θ. The only reason for using the constant $\kappa>0$ in the definition of the process $Z^{(W)}$ is to make its sample paths strictly positive, which allows one to avoid complications in the definition of the likelihood (1). The constant κ can be taken to be arbitrarily small. Note that the process W can be viewed as a map with values in the Banach space $\ell^\infty(\mathcal{X})$. In applications, sample paths of W typically possess some smoothness properties and W can also be viewed as a map taking values in a Banach space $(\mathbb{B},\|\cdot$

$\|_\infty$) for some $\mathbb{B} \subset \ell^\infty(\mathcal{X})$. We will assume that this map is Borel-measurable, so that W is a $\mathbb{B}$-valued random element. By Lemma 5.1 in van der Vaart and van Zanten (2008b), the support of W, i.e. the smallest closed set $\mathbb{B}_0 \subset \mathbb{B}$, such that $\widetilde{\mathbb{P}}(W \in \mathbb{B}_0) = 1$, is the closure in $\mathbb{B}$ of the reproducing kernel Hilbert space (RKHS) $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ attached to W. It can be shown that this RKHS can be identified with the completion of the set of maps

$$x \mapsto \sum_{i=1}^{k} \alpha_i K(y_i, x) = \widetilde{\mathbb{E}}\,[W_x H], \quad H = \sum_{i=1}^{k} \alpha_i W_{y_i},$$

under the inner product

$$\langle \widetilde{\mathbb{E}}\,[W.H_1], \mathbb{E}\,[W.H_2] \rangle_{\mathbb{H}} = \widetilde{\mathbb{E}}\,[H_1 H_2].$$

Here the α_i's range over $\mathbb{R}$ and k ranges over $\mathbb{N}$. The support of the process $Z^{(W)}$ can then be described through this characterisation of the support of the process W.

Remark 1. Other transformations of the process W can also be used to define the process $Z^{(W)}$. For instance, one can set $Z_x^{(W)} = g(W_x)$ for a fixed function g that is bounded away from zero and possesses suitable regularity properties. □

Let $N(\varepsilon, B, f)$ denote the minimum number of balls of radius ε needed to cover a subset B of a metric space with metric f. This is the ε-covering number of B.

Our main result is based on an application of Theorem 2.1 in Ghosal and van der Vaart (2001) (which is a slight modification of Theorem 2.1 in Ghosal et al. (2000)) and Theorem 2.1 from van der Vaart and van Zanten (2008a). These are provided below for the reader's convenience in an adapted form.

Theorem 1 (Ghosal and van der Vaart (2001)). *Suppose that for positive sequences* $\overline{\varepsilon}_n, \widetilde{\varepsilon}_n \to 0$, *such that* $n \min(\overline{\varepsilon}_n^2, \widetilde{\varepsilon}_n^2) \to \infty$, *constants* $c_1, c_2, c_3, c_4 > 0$ *and sets* $\Theta_n \subset \Theta$, *we have*

$$\log N(\overline{\varepsilon}_n, \Theta_n, \rho) \leq c_1 n \overline{\varepsilon}_n^2, \tag{3}$$

$$\Pi(\Theta \setminus \Theta_n) \leq c_3 e^{-n\widetilde{\varepsilon}_n^2 (c_2+4)}, \tag{4}$$

$$\Pi\left(z^{(w)} \in \Theta : \mathrm{KL}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2, \mathrm{V}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2\right) \geq c_4 e^{-c_2 n \widetilde{\varepsilon}_n^2}. \tag{5}$$

Then, for $\varepsilon_n = \max(\overline{\varepsilon}_n, \widetilde{\varepsilon}_n)$ *and a large enough constant* $M > 0$, *we have that*

$$\Pi(z^{(w)} \in \Theta : \rho(\lambda_0, z^{(w)}) \geq M\varepsilon_n | X^{(n)}) \to 0 \tag{6}$$

in $\mathbb{P}_{\lambda_0}^{(n)}$*-probability.*

Remark 2. Note that the posterior contraction rate ε_n from Theorem 1 is not uniquely defined. If ε_n is a posterior contraction rate, then so is, for instance, $(3 - \sin(n))\varepsilon_n$ as well, or in fact any sequence that converges to zero at a slower rate than ε_n. In general we are interested in finding the 'fastest' posterior contraction rate ε_n, in the sense that (6) holds for this ε_n and there is no other sequence $\varepsilon_n' \to 0$, such that $\lim_{n\to\infty} \varepsilon_n'/\varepsilon_n = 0$, for which (6) still holds with ε_n replaced by ε_n' (and perhaps the constant M replaced by another constant M'). □

The conditions of Theorem 1 merit some discussion. We restrict ourselves to heuristic reasoning only: an in-depth discussion can be found in Ghosal et al. (2000). The important conditions of the theorem are (3) and (5). Since the covering number can be thought of as measuring the size of the model, condition (3) says that in order to have posterior contraction rate ε_n, the model should not be too big. Furthermore, condition (5) tells us that in order to have the posterior contraction rate ε_n, the prior Π should put some minimal mass in the Kullback-Leibler type neighbourhoods of λ_0. Finally, condition (4) adds some additional flexibility: for our purposes it is enough to be understood in the sense that the sets Θ_n are almost the support of the prior. This condition often allows one to avoid too stringent assumptions on the parameter set Θ, such as, for instance, its compactness.

Next we need to find effective means for checking the fact that our model and the prior satisfy the conditions of Theorem 1. To that end we will employ Theorem 2.1 from van der Vaart and van Zanten (2008a). The following concept is needed in its statement: for a function $\overline{\lambda}_0 : \mathcal{X} \to \mathbb{R}$ define the function $\phi_{\overline{\lambda}_0} : \mathbb{R} \to \mathbb{R}$ through

$$\phi_{\overline{\lambda}_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-\overline{\lambda}_0\|<\varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \widetilde{\mathbb{P}}(\|W\|_\infty < \varepsilon). \tag{7}$$

This is called the concentration function of the Gaussian process W.

Theorem 2 (van der Vaart and van Zanten (2008a))**.** *Let $\overline{\lambda}_0$ be contained in the support of W. For any sequence of positive numbers $\hat{\varepsilon}_n > 0$ satisfying*

$$\phi_{\overline{\lambda}_0}(\hat{\varepsilon}_n) \leq n\hat{\varepsilon}_n^2 \tag{8}$$

and any constant $C > 1$ with $\exp(-Cn\hat{\varepsilon}_n^2) < 1/2$, there exist measurable sets $B_n \subset \mathbb{B}$, such that

$$\log N(3\hat{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq 6Cn\hat{\varepsilon}_n^2, \tag{9}$$

$$\widetilde{\mathbb{P}}(W \notin B_n) \leq e^{-Cn\hat{\varepsilon}_n^2}, \tag{10}$$

$$\widetilde{\mathbb{P}}(\|W - \overline{\lambda}_0\|_\infty < 2\hat{\varepsilon}_n) \geq e^{-n\hat{\varepsilon}_n^2}. \tag{11}$$

Comparing the three conditions (3)–(5) from Theorem 1 to the three conditions (9)–(11) from Theorem 2, we see that they are of a similar type. Once we bridge the Hellinger distance, the Kullback-Leibler divergence and the divergence V appearing in Theorem 1 with the $\|\cdot\|_\infty$-distance, Theorems 1 and 2 will yield the posterior contraction rate.

The following lemma serves the purpose of bounding the divergences appearing in Theorem 1. Its proof is found in Appendix A.

Lemma 1. *Let $\lambda_1(x) = \kappa + |w_x|$ and $\lambda_2(x) = \kappa + |v_x|$ for $w, v \in \ell^\infty(\mathcal{X})$. Then*

(i) $h(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}) \leq \frac{1}{\sqrt{\kappa}}\|w - v\|_\infty$;

(ii) $\mathrm{KL}(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}) \leq \frac{1}{\kappa}\|w - v\|_\infty^2$;

(iii) $\mathrm{V}(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}) \leq \frac{1}{\kappa}\|w - v\|_\infty^2 \left(1 + \frac{1}{\kappa}\|w - v\|_\infty\right)$.

The following is our main result.

Theorem 3. *Let $\lambda_0 = \kappa + \overline{\lambda}_0$ for $\overline{\lambda}_0 \geq 0$ that is contained in the support of W. Suppose the prior Π is the law of the process $Z^{(W)} = (Z_x^{(W)})_{x\in\mathcal{X}}$ for $Z_x^{(W)} = \kappa + |W_x|$. Then for a sequence $\varepsilon_n = \hat{\varepsilon}_n$ satisfying the assumptions of Theorem 2 and a sufficiently large constant $M > 0$, the posterior distribution for λ_0 relative to the prior Π satisfies*

$$\Pi(z^{(w)} \in \Theta : \rho(\lambda_0, z^{(w)}) > M\varepsilon_n | X^{(n)}) \to 0$$

in $\mathbb{P}_{\lambda_0}^{(n)}$-probability.

Proof. For B_n as in Theorem 2, set $\Theta_n = \{z^{(w)} : w \in B_n\}$. We need to verify the conditions of Theorem 1. Denote $c_\kappa = (1/\kappa + 1/\kappa^2)$ and let a constant $C > 1$ from Theorem 2 be large enough, so that

$$\frac{1}{4c_\kappa} \leq \frac{C}{4c_\kappa} - 4.$$

Let $\hat{\varepsilon}_n$ be a sequence of positive numbers satisfying the conditions of Theorem 2. Take $\overline{\varepsilon}_n = 3\kappa^{-1/2}\hat{\varepsilon}_n$. By Lemma 1 (i) and by inequality (9),

$$\log N(\overline{\varepsilon}_n, \Theta_n, \rho) \leq \log N(3\hat{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq 6Cn\hat{\varepsilon}_n^2 = \frac{2\kappa C}{3} n\overline{\varepsilon}_n^2,$$

which verifies (3) for the constant $c_1 = 2\kappa C/3$. Furthermore, for n large enough, so that $\widetilde{\varepsilon}_n$ is small, Lemma 1 (ii)–(iii) yields that

$$\left\{z^{(w)} \in \Theta : \mathrm{KL}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2, \mathrm{V}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2\right\} \supset \{w : c_\kappa \|w - \overline{\lambda}_0\|_\infty^2 \leq \widetilde{\varepsilon}_n^2\}.$$

Set $\widetilde{\varepsilon}_n = 2\sqrt{c_\kappa}\hat{\varepsilon}_n$. It follows from the above display and (11) that

$$\begin{aligned}\Pi\left(z^{(w)} \in \Theta : \mathrm{KL}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2, \mathrm{V}(\lambda_0, z^{(w)}) \leq \widetilde{\varepsilon}_n^2\right) &\geq \widetilde{\mathbb{P}}(\|W - \overline{\lambda}_0\|_\infty < 2\hat{\varepsilon}_n) \\ &\geq e^{-n\hat{\varepsilon}_n^2} \\ &= \exp\left(-\frac{1}{4c_\kappa} n\widetilde{\varepsilon}_n^2\right).\end{aligned}$$

This verifies (5) for $c_4 = 1$ and $c_2 \geq 1/(4c_\kappa)$. Finally, by (10),

$$\Pi(\Theta \setminus \Theta_n) = \widetilde{\mathbb{P}}(W \notin B_n) \leq e^{-Cn\hat{\varepsilon}_n^2} = \exp\left(-\frac{C}{4c_\kappa} n\widetilde{\varepsilon}_n^2\right).$$

This verifies (4) for $c_3 = 1$ and $c_2 \leq C/(4c_\kappa) - 4$. Theorem 1 then yields the posterior contraction rate $\varepsilon_n = \max(\overline{\varepsilon}_n, \widetilde{\varepsilon}_n)$. Since both $\overline{\varepsilon}_n$ and $\widetilde{\varepsilon}_n$ are proportional to $\hat{\varepsilon}_n$, we can simply take $\varepsilon_n = \hat{\varepsilon}_n$ and absorb the constants in the constant M in the statement of Theorem 1. This completes the proof. □

Remark 3. Due to boundary bias problems characteristic of kernel-type estimators, in Kutoyants (1998) the properties of a kernel estimator of $\lambda_0(x)$ are studied only for x restricted to a compact set strictly contained in $\mathcal{X}$. On the other hand, our non-parametric Bayesian approach does not suffer from this limitation (the pseudo-distance $\rho(\lambda_1, \lambda_2)$ is a global distance using all the values of λ_1 and λ_2 on $\mathcal{X}$). □

Remark 4. Motivated by applications of the so-called log-Gaussian Cox processes (see e.g. Chapter 6 in Kingman (1993) for more information on Cox processes), one could have argued that a reasonable prior Π for λ_0 would have been the process $Z^{(W)} = \left(Z_x^{(W)}\right)_{x \in \mathcal{X}}$ defined through

$$Z_x^{(W)} = e^{W_x}, \quad x \in \mathcal{X}.$$

This transforms W into a strictly positive process $Z^{(W)}$. Moreover, if the sample paths of W are, say, β-Hölder-regular, so will be the sample paths of $Z^{(W)}$. However, examination of the proof of Lemma 1 shows that in this case there does not seem to exist a good way to control the probability divergences in the statement of Theorem 1 in terms of the $\|\cdot\|_\infty$-distance. This then does not permit to invoke Theorem 2 in order to derive the posterior contraction rate. We suspect that for such a prior the posterior contracts at a suboptimal rate (in a sense that there exists some other prior, for which the posterior contracts at a faster rate ϵ_n'; cf. Remark 2). On a practical level, an objection that can be advanced against such a prior for λ_0 is that the process $Z^{(W)}$ grows too 'fast', since so does the exponential function. □

Remark 5. An interesting statistical problem related to the one we are considering in this note is non-parametric estimation of the intensity function of a cyclic Poisson point processes over $\mathcal{X} = [0,T]^d$ (i.e. a Poisson point process with a periodic intensity function). A recent reference dealing with estimation of the unknown period in this model is Belitser et al. (2012). □

3. Example of the prior

In this section we consider a concrete example of the prior and compute the posterior contraction rate for it explicitly. Let for simplicity $d=1$. We recall the definition of a β-Hölder-regular function: a function $\lambda : \mathcal{X} \to \mathbb{R}$ is said to be β-Hölder regular for $\beta > 0$, if it is continuously differentiable up to order $\lfloor \beta \rfloor$ (here $\lfloor \beta \rfloor$ denotes the largest integer strictly smaller than β. For $\lfloor \beta \rfloor = 0$ we assume that λ is continuous) and the derivative $\lambda^{(\lfloor \beta \rfloor)}$ satisfies the Hölder condition of order $\beta - \lfloor \beta \rfloor$. We will denote the space of β-Hölder-regular functions by $\mathcal{C}^\beta(\mathcal{X})$. Furthermore, $\mathcal{C}(\mathcal{X})$ will denote the space of continuous functions on $\mathcal{X}$ equipped with the uniform norm.

Example 1. Let $\overline{W} = (\overline{W}_x)_{x\in\mathcal{X}}$ be a standard Brownian motion over the time interval $\mathcal{X} = [0,1]$ and let $\eta_0, \eta_1, \ldots, \eta_{\lfloor \beta \rfloor + 1}$ be standard normal random variables. Assume that $\eta_0, \eta_1, \ldots, \eta_{\lfloor \beta \rfloor + 1}, \overline{W}$ are independent. The modified Riemann-Liouville process $W = (W_x)_{x\in\mathcal{X}}$ with Hurst parameter $\beta > 0$ is defined as

$$W_x = \sum_{k=0}^{\lfloor \beta \rfloor + 1} \eta_k x^k + \int_0^x (x-y)^{\beta - 1/2} \mathrm{d}\overline{W}_y, \quad y \in \mathcal{X},$$

see Section 4.2 in van der Vaart and van Zanten (2008a). Our prior Π will be the law of the process $Z^{(W)} = \left(Z_x^{(W)} \right)_{x\in\mathcal{X}}$ defined by (2). By Theorem 4.3 in van der Vaart and van Zanten (2008a), the support of W is the whole space $\mathcal{C}(\mathcal{X})$, and if $\lambda_0 = \kappa + \overline{\lambda}_0$ for a non-negative $\overline{\lambda}_0 \in \mathcal{C}^\beta(\mathcal{X})$, then $\phi_{\overline{\lambda}_0}(\varepsilon) \asymp \varepsilon^{-1/\beta}$ as $\varepsilon \downarrow 0$. It then follows from Theorem 3 by solving inequality (8) (cf. van der Vaart and van Zanten (2008a), pp. 1449–1450) that the posterior contracts at the rate $n^{-\beta/(2\beta+1)}$. This is the minimax estimation rate for a β-Hölder-regular function in a variety of non-parametric estimation problems. See in particular Theorem 6.5 in Kutoyants (1998) for the Poisson point processes setting. The rate $n^{-\beta/(2\beta+1)}$ can thus be thought of as an optimal posterior contraction rate in this particular setting. □

Appendix A.

Proof of Lemma 1. Part (i) follows from part (ii) and the well-known inequality

$$h^2(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2}) \leq \mathrm{KL}(\mathbb{P}_{\lambda_1}, \mathbb{P}_{\lambda_2})$$

between the squared Hellinger distance and the Kullback-Leibler divergence (alternatively, see Lemma 1.5 in Kutoyants (1998)).

We prove part (ii). Using Theorem 1.3 and Lemma 1.1 from Kutoyants (1998), we have

$$\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})=\int_{\mathcal{X}}\lambda_1(x)\log\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\mathrm{d}x-\int_{\mathcal{X}}\left\{\frac{\lambda_1(x)}{\lambda_2(x)}-1\right\}\lambda_2(x)\mathrm{d}x. \tag{12}$$

Now since $\log(1+x)\leq x$ for $x>-1$, we get that

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})&\leq\int_{\mathcal{X}}[\lambda_1(x)-\lambda_2(x)]^2\frac{1}{\lambda_2(x)}\mathrm{d}x\\&\leq\frac{1}{\kappa}\int_{\mathcal{X}}[\lambda_1(x)-\lambda_2(x)]^2\mathrm{d}x\\&\leq\frac{1}{\kappa}\|\lambda_1-\lambda_2\|_\infty^2\\&\leq\frac{1}{\kappa}\|w-v\|_\infty^2,\end{aligned}$$

where the last inequality follows from the inequality $||a|-|b||\leq|a-b|$ valid for $a,b\in\mathbb{R}$. This proves part (ii). Here we also see the role of constant $\kappa>0$.

We prove part (iii). Letting $U\sim\mathbb{P}_{\lambda_1}$ and denoting by $\mathbb{E}_{\lambda_1}[\cdot]$ the expectation under $\mathbb{P}_{\lambda_1}$, we have

$$\mathrm{V}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})=\mathbb{E}_{\lambda_1}\left[\log^2\left(\frac{\mathrm{d}\mathbb{P}_{\lambda_1}}{\mathrm{d}\mathbb{P}_{\lambda_1}}(U)\right)\right]-\mathrm{KL}^2(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2}). \tag{13}$$

Using Theorem 1.3 and Lemma 1.1 from Kutoyants (1998), as well as formula (12) above, after some uninspiring computations we get from (13) that

$$\begin{aligned}\mathrm{V}(\mathbb{P}_{\lambda_1},\mathbb{P}_{\lambda_2})&=\int_{\mathcal{X}}\lambda_1(x)\log^2\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\mathrm{d}x\\&=\int_{\lambda_1<\lambda_2}\lambda_1(x)\log^2\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\mathrm{d}x\\&+\int_{\lambda_1>\lambda_2}\lambda_1(x)\log^2\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\mathrm{d}x\\&=\mathrm{I}_1+\mathrm{I}_2,\end{aligned}$$

with an obvious definition of I_1 and I_2. Recall the elementary inequality

$$\frac{x}{1+x}\leq\log(1+x)\leq x,\quad x>-1.$$

This inequality gives that on the set $\{\lambda_1<\lambda_2\}$,

$$\log^2\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\leq\frac{1}{\lambda_1^2(x)}[\lambda_1(x)-\lambda_2(x)]^2.$$

Hence

$$\mathrm{I}_1\leq\frac{1}{\kappa}\|\lambda_1-\lambda_2\|_\infty^2\leq\frac{1}{\kappa}\|w-v\|_\infty^2.$$

On the other hand, on the set $\{\lambda_1>\lambda_2\}$,

$$\log^2\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right)\leq\left(\frac{\lambda_1(x)}{\lambda_2(x)}-1\right)^2.$$

Therefore,

$$\mathrm{I}_2\leq\int_{\lambda_1>\lambda_2}[\lambda_1(x)-\lambda_2(x)]^2\frac{\lambda_1(x)}{\lambda_2^2(x)}\mathrm{d}x$$

$$
\begin{aligned}
&= \int_{\lambda_1 > \lambda_2} [\lambda_1(x) - \lambda_2(x)]^3 \frac{1}{\lambda_2^2(x)} \mathrm{d}x + \int_{\lambda_1 > \lambda_2} [\lambda_1(x) - \lambda_2(x)]^2 \frac{1}{\lambda_2(x)} \mathrm{d}x \\
&\leq \frac{1}{\kappa^2} \|\lambda_1 - \lambda_2\|_\infty^3 + \frac{1}{\kappa} \|\lambda_1 - \lambda_2\|_\infty^2 \\
&\leq \frac{1}{\kappa} \|w - v\|_\infty^2 \left(1 + \frac{1}{\kappa} \|w - v\|_\infty \right).
\end{aligned}
$$

This completes the proof of part (iii) and hence of the lemma too. □

References

E. Belitser, P. Serra and H. van Zanten. Estimating the period of a cyclic non-homogeneous Poisson process. *Scand. J. Stat.*, doi:10.1111/j.1467-9469.2012.00806.x, 2012.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. With a discussion and a rejoinder by the authors. *Ann. Statist.*, 14:1–67, 1986.

S. Ghosal, J.K. Ghosh and A.W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28:500–531, 2000.

S. Ghosal and A.W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29:1233–1263, 2001.

J. Heikkinen and E. Arjas. Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scand. J. Statist.*, 25:435-450.

J.F.C. Kingman. *Poisson Processes.* Oxford Studies in Probability, 3. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1993.

Yu.A. Kutoyants. *Statistical Inference for Spatial Poisson Processes.* Lecture Notes in Statistics, 134. Springer-Verlag, New York, 1998.

J. Møller and R.P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes.* Monographs on Statistics and Applied Probability, 100. Chapman & Hall/CRC, Boca Raton, FL, 2004.

J. Møller and R.P. Waagepetersen. Modern statistics for spatial point processes. *Scand. J. Statist.*, 34:643–684, 2007.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

A.W. van der Vaart and J.H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36:1435–1463, 2008a.

A.W. van der Vaart and J.H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 200–222, Inst. Math. Stat. Collect., 3. Inst. Math. Statist., Beachwood, OH, 2008b.

Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

E-mail address: shota.gugushvili@math.leidenuniv.nl

Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, PO Box 94248, 1090 GE Amsterdam, The Netherlands

E-mail address: spreij@uva.nl